Controllability and Observability Trojan Detection Summary

Brian Weber

1 Abstract

This paper introduces a highly effective, low complexity method for detecting harware trojans in gate level netlists. Other previous methods were of exponential complexity, occasionally gave false positive and false negative results, and/or had reliable methods of avoiding detection. COTD is of linear complexity, has no currently known exploits, and did not throw any false positive or false negatives. In addition, COTD does not need any reference circuits to base its analysis off of. Because of it's high accuracy, every input and output signal of the trojan is detected, allowing the trojan to be easily isolated from the rest of the circuit. Due to the low complexity of COTD and because it uses tools which are already commonly used, it can easily be integrated into current testing workflows.

2 Methods

In order to detect trojans, COTD analyzes the controllability and observability (CC and CO) of every signal in the circuit. Experimentation reveals that signals with low testability (signals with high $| < CC, CO > | = \sqrt{CC^2 + CO^2}$) are much less likely to produce a wrong output during circuit authentication. Therefore, hardware trojans will likely also have a high $| < CC, CO > |$ so they are not detected during circuit authentication. COTD uses this fact to distinguish signals in trojans from valid signals.

After *CC* and *CO* are computed for every signal, they are separated into 3 groups using a machine learning tequnique called unsupervised cluster analysis. This tequnique separates a set of data into groups based on inter and intra cluster distances. It tries to form these groups such that the inter cluster distance is at a maximum, and the intra cluster distance is minimized. The three groups that are formed are signals with high *CO* values, signals with high *CC* values (both malicious signals), and signals with comparatively lower values for both *CC* and *CO* (legitimate signals).

After testing this approach on several benchmarks, it was revealed that hardware trojans in fact do have significantly different *CC* and *CO* values, so much so that they were able to be identified with 100% accuracy. There was one malicious signal in one benchmark however in wb-conmax-T100 which had a signantly lower *CC* and *CO* value than the rest of the malicious signals. Though it was correctly identified, it was still much closer to legitimate signals than the rest of the malicious signals. Taking note of this, authentication tests were run with several thousand random inputs, in which several of those inputs triggered a wrong output, indicating that even if a malicious signal is close to the *CC* and *CO* values of legitimate signals, it will be detected by authentication tests with a high probability.

3 Discussion

According to what's presented in this paper, I honestly can't think of any cons to the COTD method. It seems to be significantly stronger than any other method in almost every respect. It is much faster, and much more reliable by a wide margin. After some quick thought, it seems

like anything that can be done to have a trojan be undetected by COTD will easily be detected by other methods during the authentication of a circuit.

4 Conclusion

One thing that could be done to strengthen the claims in the paper is to run more experiments on more trojans. I was slightly disappointed by the small sample size that it was tested on. Especially on trojans that contain signals that toe the line between being placed in the malicious or legitmate pile(like wb-conmax-T100). If a very large percentage of these signals can be properly detected by using random inputs or COTD, then it would solidify COTD as a very powerful method for detecting hardware trojans. This was a very interesting and exciting paper to read.